

Mechanical response of random heteropolymers

Phillip L. Geissler and Eugene I. Shakhnovich

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

We present an analytical theory for heteropolymer deformation, as exemplified experimentally by stretching of single protein molecules. Using a mean-field replica theory, we determine phase diagrams for stress-induced unfolding of typical random sequences. This transition is sharp in the limit of infinitely long chain molecules. But for chain lengths relevant to biological macromolecules, partially unfolded conformations prevail over an intermediate range of stress. These necklace-like structures, comprised of alternating compact and extended subunits, are stabilized by quenched variations in the composition of finite chain segments. The most stable arrangements of these subunits are largely determined by preferential extension of segments rich in solvophilic monomers. This predicted significance of necklace structures explains recent observations in protein stretching experiments. We examine the statistical features of select sequences that give rise to mechanical strength and may thus have guided the evolution of proteins that carry out mechanical functions in living cells.

I. INTRODUCTION

Several recent experiments have highlighted the mechanical strength of proteins involved in muscle elasticity and cell adhesion [1–3]. When pulled from the ends, these molecules can withstand significant stress before their constituent domains unfold from compact native states to extended coil-like structures. This stress-induced unfolding occurs sharply, with threshold forces f_t on the order of 100 pN. In natural units for these systems, $f_t \sim 100k_B T/a$, where k_B is Boltzmann’s constant (which we subsequently set to unity), T is temperature, and $a \simeq 1$ nm is the size of a typical amino acid. By contrast, studies of proteins whose functions are not mechanical in nature have revealed lower threshold forces ($f_t \sim 10k_B T/a$) and less dramatic stretching behavior [4,5]. Specifically, relative fluctuations in restoring force are considerably larger than for mechanical proteins, and the stretching transition is less sharply defined. Evolution therefore appears to have designed certain proteins to unfold reproducibly under critical stress.

Results of computer simulations have lent details to this notion of mechanical design. Random heteropolymers on a lattice, which may be viewed as coarse-grained caricatures of proteins, exhibit stretching behavior that depends strongly on the sequence of constituent monomers [6]. Typical random sequences elongate smoothly under stress, passing through one or more long-lived, partially extended structures. Sequences selected for their ability to fold rapidly in the absence of stress, however, undergo a relatively sharp force-induced transition. Folding efficiency is thus correlated with mechanical strength to some extent. But reaction coordinates for protein folding are only loosely coupled to the simple mechanical variables manipulated in stretching experiments [7]. In an ensemble of fast-folding sequences, a range of mechanical stabilities is therefore expected, due to variations in sequence properties that do not affect folding dynamics. Native state topology may be one such property, according to results of protein stretching

simulations with atomistic detail [8,9]. But due to the computational expense of such simulations, they provide only anecdotal insight into the relationship between sequence and mechanical strength.

A general understanding of heteropolymer deformation can only be obtained by considering the ensemble of possible sequences. We have recently described the results of an analytical theory that takes the diversity of this ensemble properly into account [10]. Specifically, the free energies of various conformational states are averaged over the distribution of sequences, using the replica trick [11]. In this way, the mechanical response typical of random heteropolymers is determined. This article presents our theoretical approach in detail.

Because we focus on equilibrium free energetics, our theory directly applies only to *reversible* stretching, i.e., pulling rates that are much slower than rates of spontaneous unfolding. Stretching experiments, on the other hand, have been performed irreversibly, as evidenced by wide hysteresis [1]. Relating theoretical results to these experiments is made possible by an identity for nonequilibrium dynamics obtained by Jarzynski [12] and by Crooks [13]. In particular, Hummer and Szabo have shown that reversible stretching behavior may be extracted from repeated nonequilibrium measurements [14]. Equilibrium results determined in this way differ only in details from their nonequilibrium counterparts. For example, the threshold force required to unfold a mechanical protein reversibly is smaller than the corresponding nonequilibrium value, but the induced transition remains sharp [14]. Qualitative predictions of our theory are thus relevant to current experiments. Direct comparisons are possible in principle when experimental measurements have been repeated sufficiently.

The coarse features of heteropolymer response do not differ significantly from those of a homopolymer. In poor solvent ($T < \theta$), a chain molecule is transformed by stress from collapsed globule to expanded coil. This transformation occurs abruptly for homopolymers, as determined by scaling analysis [15]. Globule deformation is strongly

resisted by the cost of enlarging the polymer-solvent interface, while a coil is quite pliable. The “phase transition” between these states is thus accompanied by a sharp change in extension. At phase coexistence, the free energetic equivalence of globule and coil gives rise to necklace-like structures, in which compact and expanded subunits alternate within the chain (as sketched in Fig. 1). In the case of homopolymers, these partially extended structures are unstable away from coexistence. A phase diagram for homopolymer stretching is constructed from this physical picture in Sec. II.

Necklace structures figure more prominently in the cases of polyelectrolytes and polyampholytes. When a significant fraction of monomers carry charge, fully compact conformations are unstable, in analogy to the Rayleigh instability of charged droplets [16]. As a result, the chain segregates into a series of smaller, tethered globules. Such necklaces are the ground states of sufficiently charged polyelectrolytes, even in the absence of stress. Stretching these molecules modifies the details of structural partitioning, reducing the sizes and numbers of compact subunits and lengthening the string-like subunits that connect them [17,18].

Necklaces play an enhanced role in heteropolymer stretching as well, although for different reasons. The quenched disorder of random sequences lends different degrees of mechanical susceptibility to different regions of the chain. Depending on the extent of this heterogeneity, necklace structures may dominate over an appreciable range of stress. In effect, the globule-coil coexistence region is broadened by disorder. In Sec. III, we determine the magnitude of this broadening by analyzing a microscopic model for heteropolymer deformation. For uncorrelated sequence statistics, we show that necklaces are stabilized over a stress interval of relative width $N^{-1/2}$, where N is the number of monomers per molecule. The effects of correlations within a sequence, also examined in Sec. III, suggest that certain statistical patterns are strongly correlated with mechanical strength. These patterns are consistent with the structures of mechanical proteins designed by evolution.

In seeking a microscopic explanation for the stretching behavior of proteins, we focus on the effects of heteropolymeric disorder. Electrostatic effects are expected to influence protein stability less strongly at physiological conditions [19,20]. We also focus on the response to external stress, rather than to strain. Although the extension of protein molecules is constrained in experiments, the flexibility of unfolded segments in these modular structures mediates the applied strain. Indeed, provided the contour lengths of unfolded regions, simple elastic models account for the measured restoring forces of modular proteins. Individual, folded domains are thus effectively subjected to uniform external stress. The model we analyze in Sec. III is tailored to these external conditions appropriate for experiments and for biological function.



FIG. 1. Schematic example of a necklace-like polymeric structure. Some regions of the chain adopt locally compact, globular conformations, while others exist in extended coil-like states.

II. HOMOPOLYMER STRETCHING

We employ simple, mean-field descriptions of the conformational states relevant to polymer deformation. For instance, the free energy of a homopolymer globule relative to that of an ideal coil,

$$\mathcal{F}_g(N) \simeq B_0 \rho N + \gamma N^{2/3}, \quad (1)$$

is dominated by the effective interactions between monomers. Here, ρ is the monomer density, and γ is the surface tension of the globule-solvent interface. The energy density of monomer attractions, $B_0 = T - \theta$, stabilizes the compact state for $T < \theta$. We focus on temperatures below the θ -region, for which the globule is highly compact, i.e., $\rho v \sim 1$, where v is the volume of a monomer. At this level of description, the contribution of stress to Eq. 1 is negligible within the regime of globule stability.

We represent the coil state by a freely jointed chain with segment length a . This model provides the simplest description of polymer flexibility that ensures a finite maximum chain extension, Na . This condition is important at low temperatures, where the coil is highly extended under stress. The free energy of coil deformation is easily computed from this model [21]:

$$\mathcal{F}_c(N) = -NT \ln [y(fa/T)], \quad (2)$$

where $y(x) = \sinh(x)/x$. The stretching force at which this extended coil coexists with the compact state is determined by equating \mathcal{F}_c and \mathcal{F}_g :

$$f_t = y^{-1} \left(\exp [-(B + \gamma N^{-1/3})/T] \right) T/a. \quad (3)$$

This phase boundary is plotted as a function of temperature in Fig. 2 for various N . Qualitative features of these phase diagrams for homopolymer stretching compare well with results of lattice polymer simulations [6]. At low temperatures, a reentrant coil phase appears. This rod-like state involves small fluctuations about a fully extended structure, as has been noted in simulation work [22]. Similar reentrance has been predicted for the mechanical unzipping of DNA at low temperatures [23,24].

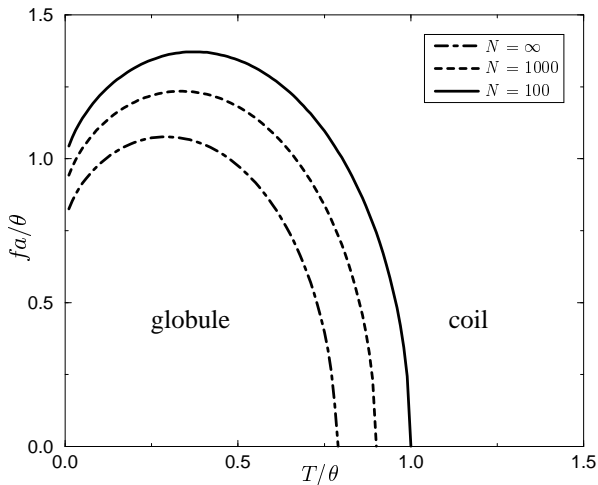


FIG. 2. Phase diagram for homopolymer stretching computed from Eq. 3. The boundary separating globule and coil states is plotted for chain lengths $N = 100$ (dot-dashed curve), $N = 1000$ (dashed curve), and for the infinite chain limit (solid curve). Here, we have taken the surface free energy density of the globule to be comparable to the energy density of monomer interactions, $\gamma \approx \theta$.

III. HETEROPOLYMER STRETCHING

Heterogeneity of monomer types has several consequences for the deformation scenario described above. First, the fully compact state is dominated by only a few distinct conformations at low temperature. The corresponding freezing transition has been analyzed thoroughly [25,26]. For necklace structures, each compact subunit can freeze in this way. Because the composition of these subunits is randomly distributed, the details of freezing will differ for each. Specifically, each subunit will have a different ground state energy, and thus a different stability. In addition, variations in sequence composition will strongly influence the solvation energetics of expanded subunits. As a result, the susceptibility of a given region of the chain to extension is effectively a random variable. This situation is illustrated in Fig. 3.

We weigh these effects of disorder using a microscopic model of random heteropolymers. For a particular realization of monomer identities, σ_i , the energy of a chain conformation is

$$\mathcal{H} = \mathcal{H}_0 + \Gamma \sum_{i \in S} \sigma_i - \mathbf{f} \cdot (\mathbf{r}_N - \mathbf{r}_1), \quad (4)$$

$$\mathcal{H}_0 = \sum_{i,j=1}^N \delta(\mathbf{r}_i - \mathbf{r}_j) (B_0 + \chi \sigma_i \sigma_j). \quad (5)$$

Here, \mathbf{r}_i denotes the position of the i th monomer in the chain, and \mathbf{f} is the external stretching force coupled to the end-to-end vector, $\mathbf{r}_N - \mathbf{r}_1$. The connectivity of consecutive monomers is implicit in Eq. 4. We take the links

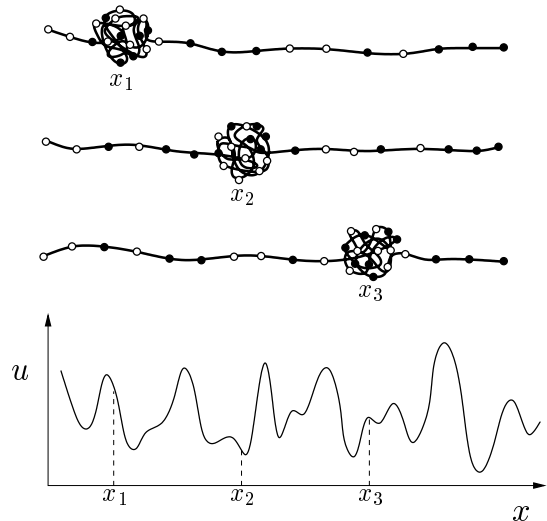


FIG. 3. Random potential $u(x)$ representing the dependence of necklace energy on the sequence location x of a globular subunit. Filled and unfilled circles denote monomers of two different types. Because the local composition of the sequence differs at x_1 , x_2 , and x_3 , globular regions at these locations will have different ground state energies. More importantly, the composition of exposed coil regions differs for the three cases. For a random sequence of monomers, the total necklace energy is thus a random variable for different subunit arrangements. The statistics of these variations determine the size of random fluctuations, Δu .

between connected monomers i and $i+1$ to be distributed according to a function $g(|\mathbf{r}_{i+1} - \mathbf{r}_i|)$ with range a . Unconnected monomers i and j interact only when they are in contact, as described by the Dirac delta function in Eq. 5. The heteropolymeric part of this interaction depends on the identities of the monomers involved. Two monomer types, $\sigma_i = \pm 1$, are possible at each point in the sequence. These types could represent, for example, amino acids with hydrophilic and hydrophobic side chains. We choose $\chi < 0$, so that each monomer attracts others of the same type most strongly. The sequence of monomer types σ_i is fixed for each realization of the heteropolymer. As in related problems, this quenched disorder requires careful treatment.

The second term in Eq. 4 describes the solvation energetics of monomers that are exposed to the external environment. The sum extends only over the set S of exposed monomers. Depending implicitly on chain conformation, this term mimics a many-body aspect of the hydrophobic effect, the tendency to bury unfavorably solvated regions of a solute. We take $\Gamma > 0$, so that monomers of type $\sigma_i = -1$ are preferentially solvated. In the following sections, we analyze the energetics of Eqs. 4 and 5 for the conformational states important to mechanical response. With these results, we construct phase diagrams for heteropolymer stretching.

A. Globule

The first term in Eq. 4, \mathcal{H}_0 , defines a model of a random copolymer in the absence of explicit solvation energetics ($\Gamma = 0$) and stretching force ($\mathbf{f} = 0$). A mean field theory for the globular state of this model was analyzed in Ref. [26] using the replica trick. At a critical temperature, T_c , the heteropolymer freezes into $O(1)$ low energy conformations. This transition is accompanied by one-step replica symmetry breaking, indicating that the hierarchy of basins of attraction in conformation space is one level deep. This result is consistent with a random energy model of the system. In other words, the density of states is well represented by drawing energies at random from a Gaussian distribution, $P(E)$, with appropriate mean, \bar{E} , and variance, Δ . Our analysis of the globular state for the model defined by Eq. 4 (with nonzero Γ and \mathbf{f}) closely follows that of Ref. [26]. We focus on the way in which solvation modifies the effective distribution of energies. As in the homopolymeric case, we neglect the effect of stretching force on globule free energetics.

In order to compute properties characteristic of an ensemble of random heteropolymers, we must average over their sequences. Because the disordered sequence is quenched, the average is correctly performed on the logarithm of the partition function, Z , rather than on Z itself. This mathematically awkward procedure can be accomplished using the replica trick [11],

$$\langle \ln Z \rangle_{\text{av}} = \lim_{n \rightarrow 0} \frac{\langle Z^n \rangle_{\text{av}} - 1}{n}, \quad (6)$$

where $\langle \dots \rangle_{\text{av}}$ denotes an average over realizations of the random sequence. For integer values of n , the quantity $\langle Z^n \rangle_{\text{av}}$ has the form of a partition function for n coupled replicas of the original system. For the model of Eq. 4,

$$\begin{aligned} \langle Z^n \rangle_{\text{av}} = & \left\langle \int \left[\prod_{\alpha=1}^n \prod_{j=1}^N d\mathbf{r}_j^\alpha g(\mathbf{r}_{j+1}^\alpha - \mathbf{r}_j^\alpha) \right] \right. \\ & \times \exp \left[-\frac{B_0}{T} \sum_{\alpha=1}^n \sum_{j=1}^N \delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha) \right] \\ & \left. \times \exp \left[-\sum_{i,j} \frac{\chi}{T} \sigma_i \delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha) \sigma_j - \frac{\Gamma}{T} \sum_{i \in S} \sigma_i \right] \right\rangle_{\text{av}}, \quad (7) \end{aligned}$$

where \mathbf{r}_i^α is the position of the i th monomer of replica number α . In order to perform the average in Eq. 7, we first rewrite the right hand side as

$$\begin{aligned} & \left\langle \int d\mathbf{r}_j^\alpha g(\mathbf{r}_{j+1}^\alpha - \mathbf{r}_j^\alpha) \exp \left[-\frac{B_0}{T} \sum_{\alpha=1}^n \sum_{j=1}^N \delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha) \right] \right. \\ & \times \exp \left[b \sum_{\alpha} \int d\mathbf{R} \sum_i \sigma_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}) \sum_j \sigma_j \delta(\mathbf{r}_j^\alpha - \mathbf{R}) \right. \\ & \left. \left. + c \sum_{\alpha} \int d\mathbf{R} \rho_{s,\alpha}(\mathbf{R}) \sum_i \sigma_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}) \right] \right\rangle_{\text{av}}, \quad (8) \end{aligned}$$

where $b = -\chi/T$, $c = -\Gamma/T$, and $\int d\mathbf{r}_j^\alpha$ denotes an integration over the monomer positions of all replicas. We have also defined

$$\rho_{s,\alpha}(\mathbf{R}) = \sum_{i \in S} \delta(\mathbf{R} - \mathbf{r}_i^\alpha) \quad (9)$$

as the density of exposed monomers at position \mathbf{R} , i.e., the spatial pattern formed by the globule boundary. Here, $\delta(\mathbf{r})$ is 1 if $\mathbf{r} = 0$, and vanishes otherwise. We perform a Hubbard-Stratonovich transformation with respect to the field

$$\sum_i \sigma_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}), \quad (10)$$

by introducing a conjugate field $\psi_\alpha(\mathbf{R})$. With this transformation, the second exponential in Eq. 8 becomes

$$\begin{aligned} & \int d\psi_\alpha(\mathbf{R}) \exp \left[-\frac{1}{4b} \sum_{\alpha} \int d\mathbf{R} \psi_\alpha^2(\mathbf{R}) \right] \\ & \times \exp \left[\sum_{\alpha} \int d\mathbf{R} \psi_\alpha(\mathbf{R}) \sum_i \sigma_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}) \right] \\ & \times \exp \left[\frac{c}{2b} \sum_{\alpha} \int d\mathbf{R} \psi_\alpha(\mathbf{R}) \rho_{s,\alpha}(\mathbf{R}) \right. \\ & \left. - \frac{c^2}{4b} \sum_{\alpha} \int d\mathbf{R} \rho_{s,\alpha}^2(\mathbf{R}) \right]. \quad (11) \end{aligned}$$

The average over sequence realizations may now be easily performed, yielding an action that is highly nonlinear in $\psi_\alpha(\mathbf{R})$. As was done in Ref. [26], we retain only the first term in a high-temperature expansion of the nonlinearity. As in that work, inclusion of additional terms does not change the leading order behavior of relevant order parameters. This simplification corresponds to treating monomer types as Gaussian, rather than binary, variables, with distribution

$$w(\sigma_i) = (2\pi\mu^2)^{-1/2} \exp(-\sigma_i^2/2\mu^2). \quad (12)$$

The variety of interaction strengths described by Eq. 12 might be appropriate for monomers that come into contact with a variety of relative orientations. It could also describe a heteropolymer with more than two possible monomer types.

Averaging over the effective distribution of monomer identities, and scaling the field $\psi_\alpha(\mathbf{R})$ by $2b$, we obtain

$$\begin{aligned}
\langle Z^n \rangle_{\text{av}} = & \left\langle \exp \left[-\frac{B_0}{T} \sum_{\alpha} \int d\mathbf{R} \rho_{\alpha}^2(\mathbf{R}) - \frac{c^2}{4b} \sum_{\alpha} \int d\mathbf{R} \rho_{s,\alpha}^2(\mathbf{R}) \right] \right. \\
& \times \int \mathcal{D}\psi_{\alpha}(\mathbf{R}) \exp \left[-b \int d\mathbf{R} \psi_{\alpha}^2(\mathbf{R}) \right. \\
& \left. \left. + 2b^2 \mu^2 \sum_{\alpha,\beta} \int d\mathbf{R}_1 d\mathbf{R}_2 \psi_{\alpha}(\mathbf{R}_1) \psi_{\beta}(\mathbf{R}_2) Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \right] \right. \\
& \left. \times \exp \left[c \sum_{\alpha} \int d\mathbf{R} \psi_{\alpha}(\mathbf{R}) \rho_{s,\alpha}(\mathbf{R}) \right] \right\rangle_{\text{th}}, \quad (13)
\end{aligned}$$

where $\langle \dots \rangle_{\text{th}}$ denotes a thermal average over the statistics of monomer links imposed by $g(\mathbf{r}_{i+1}^{\alpha} - \mathbf{r}_i^{\alpha})$. In Eq. 13, we have additionally introduced two fields: a single-replica density field, $\rho_{\alpha}(\mathbf{R} = \sum_i \delta(\mathbf{r}_i - \mathbf{R}))$, and a field describing the conformational similarity of two replicas,

$$Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \sum_i \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}_1) \delta(\mathbf{r}_i^{\beta} - \mathbf{R}_2). \quad (14)$$

Since density fluctuations are negligible in the globular state, $\rho_{\alpha}(\mathbf{R})$ may be approximately replaced by its mean value, $\rho \sim v^{-1}$, where v is the volume of a typical monomer. Similarly, the surface density, $\rho_s(\mathbf{R})$, is essentially fixed in the globular state. The replica overlap function, $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)$, however, is an important measure of the population of different conformational basins of attraction. It is thus a useful order parameter to describe freezing of random heteropolymers into their lowest energy conformations.

Because density is virtually constant within the globule, $Q_{\alpha\beta}$ is a function of $\mathbf{R}_1 - \mathbf{R}_2$ only, and its Fourier transform depends on a single wavevector \mathbf{k} :

$$\hat{Q}_{\alpha\beta}(\mathbf{k}) = \int d(\mathbf{R}_1 - \mathbf{R}_2) Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \exp[i\mathbf{k} \cdot (\mathbf{R}_1 - \mathbf{R}_2)]. \quad (15)$$

With a Fourier representation of $\psi_{\alpha}(\mathbf{R})$, the right hand side of Eq. 13 becomes

$$\begin{aligned}
& \left\langle \int \mathcal{D}\hat{\psi}_{\alpha}(\mathbf{k}) \exp \left[V \sum_{\alpha,\beta} \sum_{\mathbf{k}} P_{\alpha\beta}(\mathbf{k}) \hat{\psi}_{\alpha}(\mathbf{k}) \hat{\psi}_{\beta}(-\mathbf{k}) \right. \right. \\
& \left. \left. + Vc \sum_{\alpha} \sum_{\mathbf{k}} \hat{\rho}_s(\mathbf{k}) \hat{\psi}_{\alpha}(-\mathbf{k}) \right] \right\rangle_{\text{th}}, \quad (16)
\end{aligned}$$

where

$$P_{\alpha\beta}(\mathbf{k}) = -b\delta_{\alpha\beta} + 2\mu^2 b^2 \hat{Q}_{\alpha\beta}(\mathbf{k}). \quad (17)$$

In Eq. 16 we have omitted the first exponential of Eq. 13, which contributes an irrelevant multiplicative constant.

Following the analysis of Ref. [26], we use the replica overlap function as the order parameter for a mean field

theory of heteropolymer freezing. To this end, we rewrite Eq. 16 as a functional integral over possible realizations of $Q_{\alpha\beta}$:

$$\langle Z^n \rangle_{\text{av}} = \int \mathcal{D}\hat{Q}_{\alpha\beta}(\mathbf{k}) \exp[-E\{\hat{Q}_{\alpha\beta}(\mathbf{k})\} + S\{\hat{Q}_{\alpha\beta}(\mathbf{k})\}] \quad (18)$$

Here, E is the effective energy of a particular realization:

$$\begin{aligned}
E\{\hat{Q}_{\alpha\beta}(\mathbf{k})\} = & \ln \int \mathcal{D}\psi_{\alpha}(\mathbf{R}) \\
& \times \exp \left[V \sum_{\alpha,\beta} \sum_{\mathbf{k}} P_{\alpha\beta}(\mathbf{k}) \hat{\psi}_{\alpha}(\mathbf{k}) \hat{\psi}_{\beta}(-\mathbf{k}) \right. \\
& \left. + Vc \sum_{\alpha} \sum_{\mathbf{k}} \hat{\rho}_s(\mathbf{k}) \hat{\psi}_{\alpha}(-\mathbf{k}) \right] \quad (19)
\end{aligned}$$

$$= \int d\mathbf{k} \left[\ln \det P_{\alpha\beta}(\mathbf{k}) - \frac{c^2}{4} \sum_{\alpha,\beta} |\hat{\rho}_s(\mathbf{k})|^2 P_{\alpha\beta}^{-1}(\mathbf{k}) \right]. \quad (20)$$

Similarly, S is an effective entropy describing the number of conformations consistent with a particular realization of $Q_{\alpha\beta}$:

$$\begin{aligned}
S\{\hat{Q}_{\alpha\beta}(\mathbf{k})\} = & \ln \left\langle \delta \left(\hat{Q}_{\alpha\beta}(\mathbf{k}) - V \sum_i \exp[i\mathbf{k} \cdot (\mathbf{r}_i^{\alpha} - \mathbf{r}_i^{\beta})] \right) \right\rangle_{\text{th}}. \quad (21)
\end{aligned}$$

We will approximate the integral in Eq. 18 by optimizing the free energy with respect to replica overlap.

We imagine that a hierarchy exists for basins of attraction in conformation space, as is done in the theory of spin glasses [11]. It is then natural to sort replicas into groups, such that replicas belonging to the same group overlap most strongly. This grouping determines the structure of $Q_{\alpha\beta}$. If α and β are in the same group, $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)$ is nearly $\rho\delta(\mathbf{R}_1 - \mathbf{R}_2)$. If α and β belong to widely different groups, $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \approx 0$. The one-step replica symmetry breaking demonstrated in Ref. [26] is not altered by the solvation energetics in Eq. 4, because the scaling properties of $Q_{\alpha\beta}$ are unchanged. Consequently, overlap between replicas is binary:

$$\hat{Q}_{\alpha\beta}(\mathbf{k}) = \begin{cases} \rho, & \text{for } \alpha, \beta \text{ in the same group,} \\ 0, & \text{for } \alpha, \beta \text{ in different groups.} \end{cases} \quad (22)$$

Note that $\hat{Q}_{\alpha\beta}(\mathbf{k})$ is independent of \mathbf{k} , since replica overlap is either absent or microscopically complete. Together with the number of replicas in each group, Eq. 22 specifies $\hat{Q}_{\alpha\beta}(\mathbf{k})$ completely.

The limit $n \rightarrow 0$ in Eq. 6 is most conveniently taken using a continuous representation of the replica overlap matrix. In this limit, the “number” of replicas in a group, x_0 , lies between 0 and 1, and summations over replica

indices are replaced by integrations on the interval $[0, 1]$. The first term in Eq. 20 was computed in Ref. [26] using the continuous form of $Q_{\alpha\beta}$ introduced by Parisi:

$$\ln [\det P_{\alpha\beta}] = \ln b + \frac{\ln(1 - \gamma x_0)}{x_0}, \quad (23)$$

where $\gamma = 2b\mu^2\rho$. We evaluate the second term in Eq. 20 using identities derived in Ref. [27] for the Parisi matrix:

$$\sum_{\alpha,\beta} |\hat{\rho}_s(\mathbf{k})|^2 P_{\alpha\beta}^{-1}(\mathbf{k}) = \frac{n}{b(1 - \gamma x_0)} A, \quad (24)$$

where A is the surface area of the globule. The loss of entropy due to the grouping of replicas described by $Q_{\alpha\beta}$ was also computed in Ref. [26] as $S = NnS/x_0$. Here, $s = \ln a^3/v$ is the entropy loss per monomer of constraining a replica to correspond to other replicas in its group at a microscopic level. Combining these results, and noting that wavevector summations contribute only unimportant factors of volume, we obtain the replica free energy density:

$$\frac{\mathcal{F}(x_0)}{nN} = \ln b + \frac{\ln(1 - \gamma x_0)}{x_0} - \frac{c^2}{4b}(1 - \gamma x_0)^{-1} \frac{A}{N} - \frac{s}{x_0}. \quad (25)$$

Eq. 25 differs from the corresponding result in Ref. [26] only by the term proportional to $A/N \sim N^{-1/3}$.

We now employ a mean field approximation by optimizing the free energy density with respect to x_0 . (Because the number of pairs of replicas is negative in the limit $n \rightarrow 0$, the appropriate extremum is in fact a maximum of $F(x_0)$.) To lowest nonvanishing order in x_0 , the mean field solution is

$$x_0 = \gamma^{-1} \sqrt{\frac{2s}{1 + c^2 A/2b\gamma N}}. \quad (26)$$

From Eq. 26, we may identify the transition temperature, T_c , may be identified at which freezing occurs, i.e., at which x_0 first deviates from unity:

$$T_c = (2s)^{-1/2} (-2\chi\mu^2\rho - \frac{\Gamma^2 A}{4\chi N}) + O(N^{-2/3}). \quad (27)$$

Comparing this result with Eq. 3.9 of Ref. [26], we find that the solvation term in Eq. 4 raises T_c by an amount $\sim N^{-1/3}$.

Together with the average energy of non-native conformations, the freezing temperature in Eq. 27 determines the parameters of a random energy model corresponding to the random heteropolymer. The effective distribution of energies is given by

$$P(E) = (2\pi)^{1/2} \exp [-(E - \bar{E})^2 / N\Delta^2], \quad (28)$$

where $\Delta = \sqrt{2s}T_c$ and $\bar{E} = B_0\rho N$. This distribution is dominated by states in the interval $\bar{E} - N^{1/2}\Delta <$

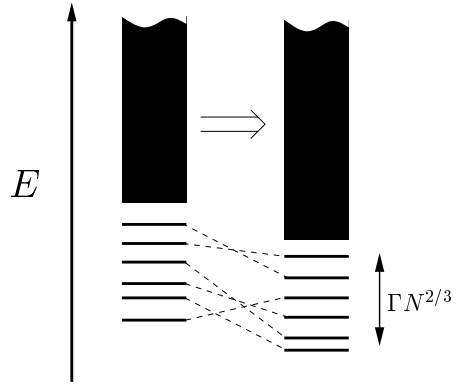


FIG. 4. Schematic effect of solvation on the distribution of globule energies. Low-lying energy levels, as well as the essentially continuous part of the energy spectrum, are sketched at left for a heteropolymer without differential solvation of monomer types. Surface energetics reorder these levels on a scale $\Gamma N^{2/3}$ (as shown at right). The basic form of the distribution is unchanged.

$E < \bar{E} + N^{1/2}\Delta$. At energies just below a critical value, $E^* = \bar{E} - N\Delta s^{1/2}$, the number of states is $O(1)$, while just above E^* the number is exponentially large. The ground states of particular random sequences are distributed narrowly about E^* [28]. Solvation thus lowers the average ground state energy by an amount $(\Gamma^2\rho/4|\chi|)s^{1/2}A$.

Solvation of the globule surface selects a ground state from the set of conformations with monomer interaction energies $E < E^* + \Gamma N^{2/3}$. This selection is illustrated in Fig. 4. If the energy scale of solvation is small, $|\Gamma| \ll |\chi|$, the shift in ground state energy will be a negligible fraction of the optimum surface energy, $\Gamma\rho A$. In this case, the set of low-energy conformations from which to select is small, and it is unlikely that one of these conformations presents a predominantly solvophilic surface. If, on the other hand, $|\Gamma| \approx |\chi|$, solvation can be an important factor in determining the ground state. In this case, there is a reasonable probability that a conformation with $E < E^* + \Gamma N^{2/3}$ has favorable solvation energy. Here, the shift in ground state energy will be comparable to $\Gamma\rho A$, and the surface of the native state will be largely solvophilic. This solvation effect does not strongly influence the freezing behavior studied in Ref. [26]. But it does represent the energetic contribution most sensitive to variations in sequence composition. It is therefore significant for our analysis of necklace structures.

B. Coil

For the coil state of a heteropolymer, we must add solvation energetics to the free energy in Eq. 2. For simplicity, we consider only random sequences whose total com-

position is fixed by $\sum_i \sigma_i = 0$. Since nearly all monomers are exposed to solvent in the expanded coil state, this constraint causes the solvation energy in Eq. 4 to vanish. Thus, the free energy of this heteropolymeric coil is identical to that of the homopolymeric coil discussed in Sec. II. In the next section, we consider necklace structures of a random heteropolymer. In these structures, short segments of the chain exist in coil-like states. The sequence composition of these exposed segments is not constrained, and the solvation energy need not vanish. We will show this to be an important stabilizing effect for necklace structures.

C. Necklace structures

For a random heteropolymer, the numbers and sizes of globular subunits do not provide an adequate description of a necklace-like structure. Due to the heterogeneity of monomer interactions, the energy of such a structure depends as well on the arrangement of subunits within the sequence. This dependence on globule location, x , can be represented as the effect of an external, random potential $u(x)$ with fluctuations of size Δu , as illustrated in Fig. 3. If Δu is large, it is likely that a particular arrangement lies much lower in energy than others. Over some range of stretching forces, it is possible that such low-energy necklace structures are preferred over pure globule and coil states. In this case, necklaces will play a significant role in the response to stretching. Our analysis of heteropolymeric necklaces is guided by this perspective. First, we characterize the statistics of the effective random potential. We then identify situations in which quenched disorder stabilizes necklace structures significantly.

Although we take sequence composition to be fixed for the chain as a whole, the composition q of local regions is distributed according to

$$p(q) \propto \exp \left[-\frac{1}{2} q^2 \left(\frac{1}{M} - \frac{1}{N} \right)^{-1} \right], \quad (29)$$

$$q = \frac{1}{M} \sum_{i \in \text{segment}} \sigma_i, \quad (30)$$

where M is the length of the segment. Because different segments of the sequence have different compositions, their local free energetics will vary. In a region with composition q , the apparent distribution of monomer types is modified from Eq. 12,

$$w(\sigma_i; q) \propto \exp [-(\sigma_i - q)^2 / 2(1 - q^2)\mu^2]. \quad (31)$$

These modified sequence statistics alter the local distribution of energies, for both globule and coil subunits.

In the context of the random energy model discussed in Sec. III.A, the variation in local sequence statistics modulates the mean and variance of conformational energies.

In effect, each segment of the sequence has a different associated random energy model. A globular subunit will therefore have a different ground state energy for each location in the sequence. Specifically, the local value of q shifts the mean energy by an amount $\chi q^2 \rho M + \Gamma q \rho A$, and reduces the variance of monomer identities, μ , by a factor $\sqrt{1 - q^2}$. As a result, the characteristic ground state energy in a sequence region with composition q is

$$E^* = B_0 \rho M + \chi q^2 \rho M + \Gamma q \rho A + \chi \mu^2 (1 - q^2) \rho M + O(M^{1/3}). \quad (32)$$

Variations in E^* along the sequence contribute to the random potential $u(x)$. The magnitude of this contribution is computed by averaging variations in E^* over the distribution of q in Eq. 29:

$$\sqrt{\langle (\delta E^*)^2 \rangle_q} \sim \Gamma \mu \rho M^{1/6}. \quad (33)$$

Fluctuations in $u(x)$ due to energetics of a globular region of length M thus arise from solvation at leading order, and grow rather slowly with increasing M .

Variations in the solvation of coil regions are more sizable. In a region of length M and composition q , the solvation energy is $\Gamma q M$. Fluctuations of this energy are of size $\Gamma \mu M^{1/2} (1 - M/N)^{-1/2}$. These variations are considerably larger than those of globule energetics for large M , and set the scale of Δu . We show below that this solvation effect is sufficient to stabilize necklace structures for long but finite chains.

We focus on necklace structures including m globular regions, each with M monomers. Thermodynamics of this class of structures may be approximated by drawing Ω values at random from a Gaussian distribution with variance Δu . Here, Ω is the number of statistically independent arrangements of the globular regions, $\Omega \approx [M^{-m} (N - M + 1)(N - 2M + 1) \dots (N - mM + 1) / m!]$. The thermodynamics of large systems with independently distributed random energies are well known [28, 29]. In this case, the free energy of spontaneous fluctuations in the random potential $u(x)$ is given by

$$\mathcal{F}_{\text{rand}}(M, m) = \begin{cases} -\ln \Omega T [1 + (T_\ell / T)^2], & T > T_\ell \\ -2 \ln \Omega T_\ell, & T \leq T_\ell. \end{cases} \quad (34)$$

In Eq. 34, $T_\ell = \Gamma \mu (mM)^{1/2} (1 - mM/N)^{1/2} / (2 \ln \Omega)^{1/2}$ is the temperature at which globular regions become localized in the sequence. For $T < T_\ell$, the necklace has a frozen arrangement of subunits, and does not reorganize. Combining Eq. 34 with Eqs. 1 and 2, we obtain the total free energy of a necklace structure,

$$\mathcal{F}_{\text{neck}}(M, m) = m \mathcal{F}_g(M) + \mathcal{F}_c(N - mM) + \mathcal{F}_{\text{rand}}(M, m). \quad (35)$$

The pure globule and coil states of a heteropolymer are also described by Eq. 35 for $M = N$ and $M = 0$, respectively.

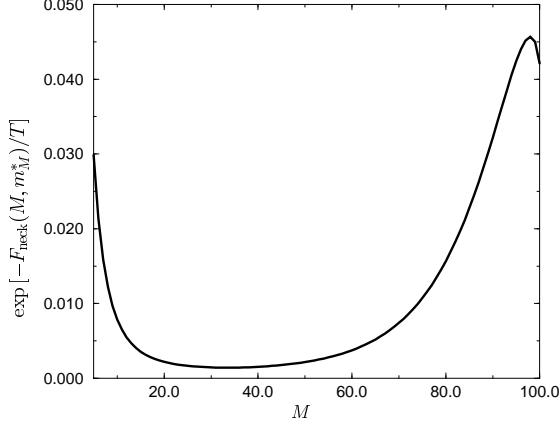


FIG. 5. Distribution of necklace structures, $p(M)$, for the thermodynamic state $T = 0.5\theta$, $f = 0.95\theta/a$ with $N = 100$. Necklaces consisting of a single large globule ($M = O(N)$) and short coil regions are most probable at these conditions. But necklaces with many small globules ($M \ll N$) are also strongly represented. Globules of intermediate size ($M \approx N/2$) occur with a small but nonnegligible probability. A minimum size of 5 monomers is chosen for globular regions.

The scaling of Eq. 34 suggests that the distribution of globular region sizes may be quite broad. For a necklace with small globules ($M \ll N$), $\mathcal{F}_{\text{rand}}$ is optimized with a large number of regions, $m = O(N)$. In this case, $\mathcal{F}_{\text{rand}} \sim N^{1/2}$. For large globules, the constraint that $mM < N$ requires that $m = O(1)$. In this case as well, $\mathcal{F}_{\text{rand}} \sim N^{1/2}$. Surface effects will yield a preference for a small number of large globular regions, but this scaling indicates that a broad ensemble of globule sizes may be important at a single thermodynamic state. This possibility is demonstrated in Fig. 5, in which the distribution of globule sizes, $p(M) \propto \exp[-\sum'_m \mathcal{F}_{\text{neck}}(M, m)/T]$, is plotted as a function of M for a thermodynamic state near the globule-coil transition. Here, primed sums are restricted to $m \leq N/M$. The weight of necklaces consisting of a single large globule is comparable to that of necklaces comprised of many small globules. As a result, fluctuations in polymer extension are considerable near the transition.

We determine a phase diagram for stretching of random heteropolymers by computing the total fraction of monomers belonging to globular regions, ϕ_g :

$$\phi_g \simeq \frac{\sum_M \sum'_m m M \exp[-\mathcal{F}_{\text{neck}}(m, M)/T]}{\sum_M \sum'_m \exp[-\mathcal{F}_{\text{neck}}(m, M)/T]}. \quad (36)$$

When $\phi_g > 0.05$, we consider the heteropolymer to be in a globular state. Similarly, when $\phi_g < 0.95$, we consider it to be in a coil state. In the intermediate regime, $0.05 < \phi_g < 0.95$, necklace structures are prevalent. Results are plotted in Fig. 6. For $N = 100$, the extent of the necklace phase prevents the possibility of a sharp

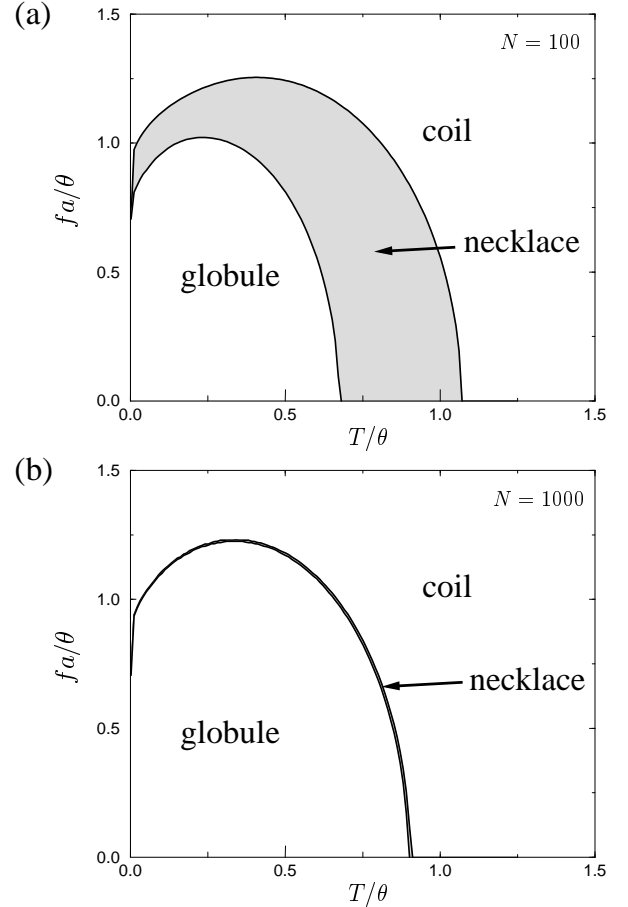


FIG. 6. Phase diagrams for random heteropolymer stretching computed from Eq. 36. Shaded regions mark a necklace phase, in which the fraction of monomers belonging to globular regions lies between 0.05 and 0.95. For these calculations, we have chosen the scale of monomer solvation energetics to coincide with that of interactions between monomers, $\Gamma \approx \theta$. We have also chosen the variance of monomer identities, μ , to be unity.

transition from globule to coil. This result is consistent with simulated stretching of a short ($N = 27$) random heteropolymer, suggesting that intermediates states described in Ref. [6] should be identified with the necklace structures we consider here. But the stabilization of necklaces arising from $\mathcal{F}_{\text{rand}}$ scales only as $N^{1/2}$, and is overcome for long chains by $O(N)$ contributions of the first two terms in Eq. 34. The range of stretching forces over which necklaces are stable is therefore significantly diminished for $N = 1000$, and vanishes as $N \rightarrow \infty$. Stretching behavior of infinitely long random heteropolymers is indistinguishable from that of homopolymers. But for chain lengths relevant to macromolecules, necklace structures can play an important role. The large fluctuations in extension accompanying these structures explains the observed stretching behavior of proteins like barnase [5] that are not designed for mechanical functions.

D. Sequence statistics

In the previous section we showed that the distribution of local sequence compositions plays an important role in the stretching of heteropolymers with uncorrelated random sequences. Introducing correlations into the sequence statistics will alter this local composition, and may therefore affect stretching behavior strongly. Here we consider three simple, prototypical forms of correlated statistics.

When monomers of the same type are likely to be found within a correlation “length” ξ in the sequence, clusters of like monomers occur with high probability. The weight of necklace structures is clearly enhanced for such blocky sequences, since regions of the chain with unfavorable solvation energy may be almost completely shielded from the solvent. Specifically, the distribution of local sequence compositions is nearly binary and independent of region size M for $M < \xi$:

$$p(q) \simeq \frac{1}{2}(\delta_{q,1} + \delta_{q,-1}). \quad (37)$$

As a result, fluctuations in solvation energy of exposed coil regions are of size $\Gamma\xi$. If $\xi \sim N$, the free energy stabilizing necklace structures is also macroscopic, $\mathcal{F}_{\text{rand}} \sim N$. With this macroscopic stabilization, the necklace phase will be stable over a finite range of f even as $N \rightarrow \infty$.

When correlations in monomer type decay algebraically, rather than exponentially, similarly dramatic fluctuations in local composition are possible. Specifically, power law correlations with decay exponent η yield $\langle(\delta q)^2\rangle_q \sim M^{-\eta}$. Because the sequence is one-dimensional, fluctuations are enhanced only for $\eta \leq 1$. At the crossover ($\eta = 1$), $\langle(\delta q)^2\rangle_q \sim M^{-1} \ln M$, providing a weak stabilization of necklaces. But as $\eta \rightarrow 0$, clusters of like monomers may be arbitrarily large. In this limit, necklace stabilization again becomes macroscopic.

Anticorrelations between like monomers, on the other hand, tends to destabilize necklace structures. In this case, the local composition is essentially “neutral” (i.e., $q \approx 0$) for regions larger than the scale of anticorrelations, ξ . For the statistics of neutrality fluctuations in Coulombic systems, $\langle(\delta q)^2\rangle_q \sim M^{-1/2}$, fluctuations in the effective random potential, $\Delta u = O(1)$, are especially small. On scales larger than ξ , the chain is effectively a homopolymer. Consequently, necklace structures are not sufficiently stable to appear away from globule-coil coexistence.

These results suggest some basic principles for designing mechanically robust heteropolymers. Typical random sequences are poor candidates, since they tend to form partially unfolded necklace structures under strain. Sequences in which solvophobic groups are heavily clustered together will typically also permit stable ensembles of necklace structures. Most promising are sequences whose correlations suppress fluctuations in local composition. These may represent, for example, molecules

with widely distributed hydrophobic groups. The compact native structures of such molecules generally include important contacts linking distant segments of the chain. It is in part this topology imposed by nonlocal contacts that provides a collective resistance to strain. Indeed, a common structural motif of mechanical proteins, β -sheet secondary structure, is typically rich in nonlocal hydrophobic contacts. The elements of sequence statistics we predict to be favorable for mechanical strength are thus related to topological features of the native state suggested by computer simulations [8,9]. It will be interesting to see how these basic principles compare with simulations of evolutionary design for mechanical strength. For commonly used, coarse-grained models of proteins, such simulations should be feasible using current computational resources and are currently underway in our laboratory.

ACKNOWLEDGMENTS

We thank A. Yu. Grosberg and D. Klimov for fruitful discussions and critical readings of this manuscript. This work was supported by NIH.

-
- [1] A. F. Oberhauser, P. E. Marszalek, H. P. Erickson, and J. M. Fernandez, *Nature* **393**, 181 (1998).
 - [2] P. E. Marszalek *et al.*, *Nature* **402**, 100 (1999).
 - [3] H. Li *et al.*, *Proc. Nat. Acad. Sci.* **97**, 6527 (2000).
 - [4] G. Yang *et al.*, *Proc. Nat. Acad. Sci.* **97**, 139 (2000).
 - [5] R. Best *et al.*, *Biophys. J.* **81**, 2344 (2001).
 - [6] D. K. Klimov and D. Thirumalai, *Proc. Nat. Acad. Sci.* **96**, 6166 (1999).
 - [7] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, *Proc. Nat. Acad. Sci.* **96**, 2031 (1999).
 - [8] H. Lu and K. Schulten, *Proteins: Struct. Func. Genet.* **35**, 453 (1999).
 - [9] E. Paci and M. Karplus, *Proc. Nat. Acad. Sci.* **97**, 6521 (2000).
 - [10] P. L. Geissler and E. I. Shakhnovich, submitted to *Phys. Rev. Lett.*
 - [11] K. Binder and A. P. Young, *Rev. Mod. Phys.* **58**, 801 (1986).
 - [12] C. Jarzynski, *Phys. Rev. E* **56**, 5018 (1997).
 - [13] G. E. Crooks, *Phys. Rev. E* **61**, 2361 (2000).
 - [14] G. Hummer and A. Szabo, *Proc. Nat. Acad. Sci.* **98**, 3658 (2001).
 - [15] A. Halperin and E. B. Zhulina, *Europhys. Lett.* **15**, 417 (1991).
 - [16] Y. Kantor and M. Kardar, *Phys. Rev. E* **51**, 1299 (1995).
 - [17] T. A. Vilgis, A. Johner, and J. F. Joanny, *Eur. Phys. J. E* **2**, 289 (2000).
 - [18] M. N. Tamashiro and H. Schiessel, *Macromolecules* **33**, 5263 (2000).

- [19] K. A. Dill, *Biochemistry* **29**, 7133 (1990).
- [20] L. Xiao and B. Honig, *J. Mol. Biol.* **289**, 1435 (1999).
- [21] A. Y. Grosberg and A. R. Khokhlov, *Statistical Mechanics of Chain Molecules* (American Institute of Physics, Woodbury, NY, 1994).
- [22] D. K. Klimov and D. Thirumalai, *J. Phys. Chem. B* **105**, 6648 (2001).
- [23] D. Marenduzzo, A. Trovato, and A. Maritan, *Phys. Rev. E* **64**, 031901 (2001).
- [24] E. A. Mukamel and E. I. Shakhnovich, *cond-mat/0108447*.
- [25] E. I. Shakhnovich and A. M. Gutin, *Biophys. Chem.* **34**, 187 (1989).
- [26] C. D. Sfatos, A. M. Gutin, and E. I. Shakhnovich, *Phys. Rev. E* **48**, 465 (1993).
- [27] M. Mezard and G. Parisi, *J. Phys. (Paris) I* **1**, 809 (1991).
- [28] B. Derrida, *Phys. Rev. B* **24**, 2613 (1981).
- [29] B. Derrida, *Phys. Rev. Lett.* **45**, 79 (1980).